Object Recognition as Machine Translation: Learning a Lexicon for a Fixed image Vocabulary

Pinar Duygulu, Kobus Barnard, Nando de Freitas and David Forsyth

UC Berkeley Digital Library Project
UBC Computer Science

Funding provided by NSF Digital Library Initiative II.
Kobus Barnard also receives funding from NSERC (Canada)
Pinar Duygulu is also supported by TUBITAK (Turkey)

Problems in Object Recognition

- What is an object ?

- How to model?

- Scale

Our Approach

Object recognition on a large scale is linking words with image regions

Use joint probability of words and pictures in large datasets

tiger grass cat

Auto-Annotating Images

Finding words for the images

→ tiger grass cat

Barnard, Forsyth (ICCV 2001) , Barnard, Duygulu, Forsyth (CVPR 2001)

Other related work : Maron 98, Mori 99

Annotation vs Recognition

?

tiger cat grass

Cannot be solved with one example

Statistical Machine Translation

Data: Aligned sentences, but word correspondences are unknown

“the beautiful sun”

“le soleil beau”

Brown, Della Pietra, Della Pietra & Mercer 93

Statistical Machine Translation

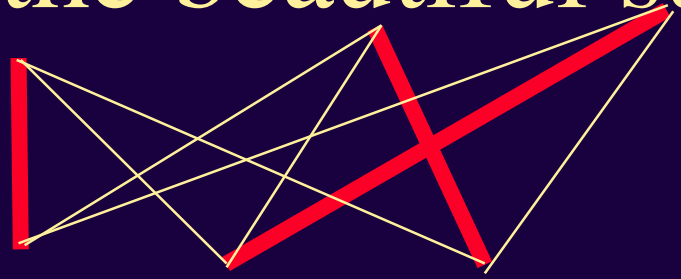
Given the correspondences, we can estimate the translation p(sun|soleil)

Given the probabilities, we can estimate the correspondences

Statistical Machine Translation

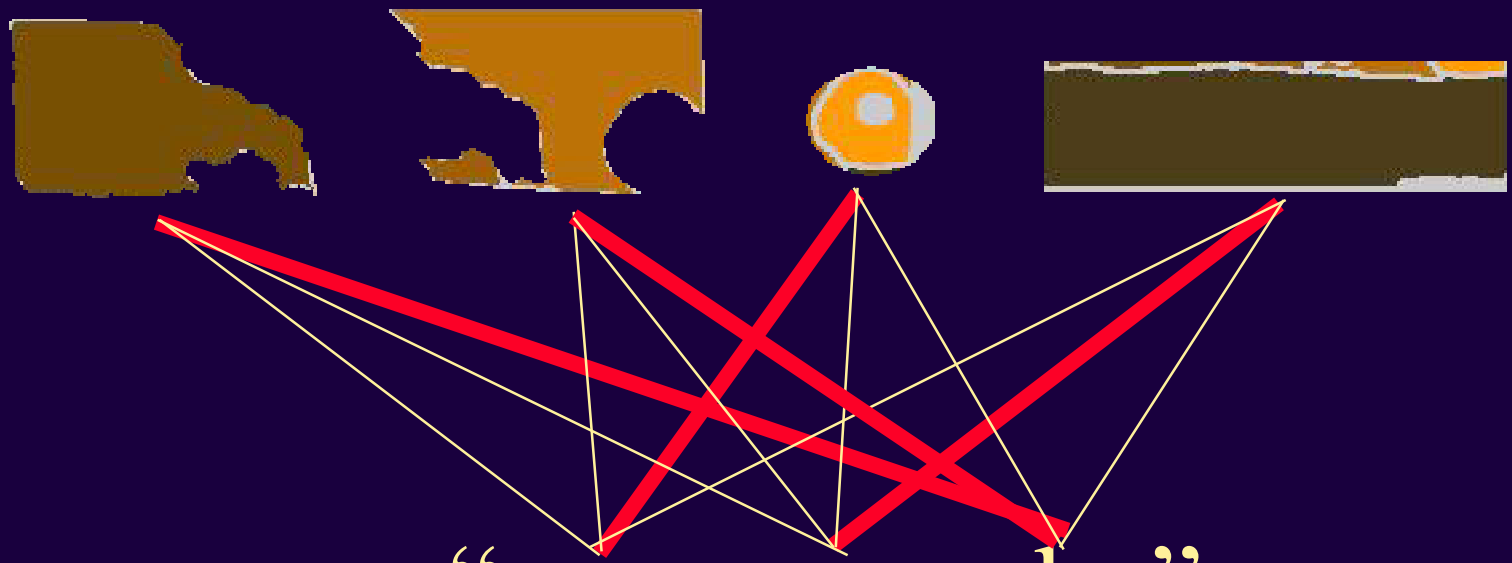
Enough data + EM, we can obtain the translation p(sun|soleil)=1

“the beautiful sun”



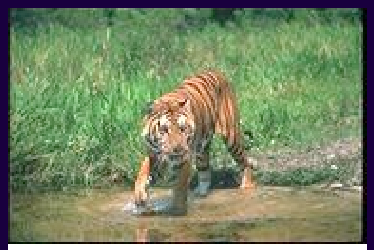
“le soleil beau”

Multimedia Translation

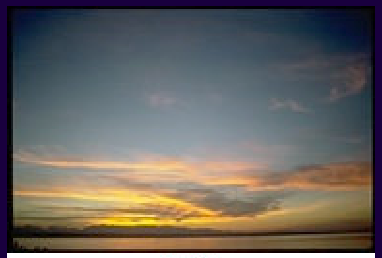


Corel Database

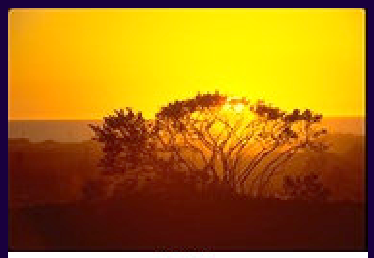
118011
WATER HARBOR
SKY CLOUDS



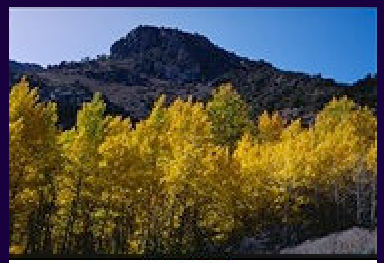
TIGER CAT WATER GRASS



1090
SUN CLOUDS
WATER SKY



1015
SUN TREE
PLAIN SKY



143078
MOUNTAINS TREES
aspens VALLEY

102042
MUSEUM memorial
FLAGS GRASS

119094
GARDEN BUILDING
FLOWERS TREES

131007
GARDEN FLOWERS
HOUSE TREES

392 CD's, each consisting of 100 annotated images.

Input

Image processing*

sun sky waves sea

Each region is described by a set of features

- Region size
- Position
- Color
- Oriented energy (12 filters)
- Simple shape features

*Thanks to Blobworld team [Carson, Belongie, Greenspan, Malik], N-cuts team [Shi, Tal, Malik]

Tokenization

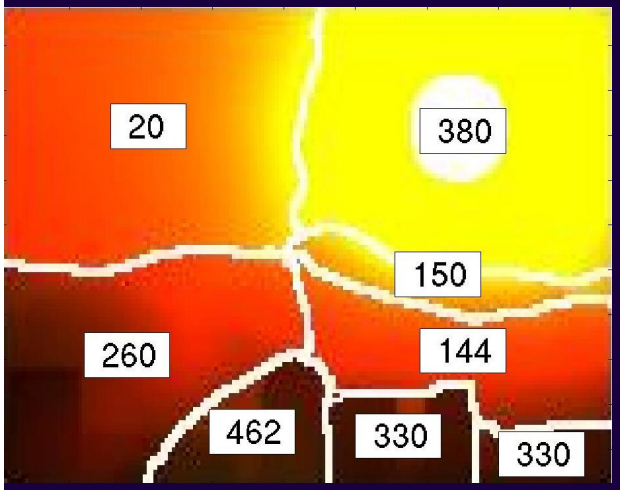
- Words → word tokens
- Image segments
  - •represented by 40 features
    (size, position, color, texture and shape)
  - •k-means to cluster features
  - •best cluster for the blob → blob tokens

Data

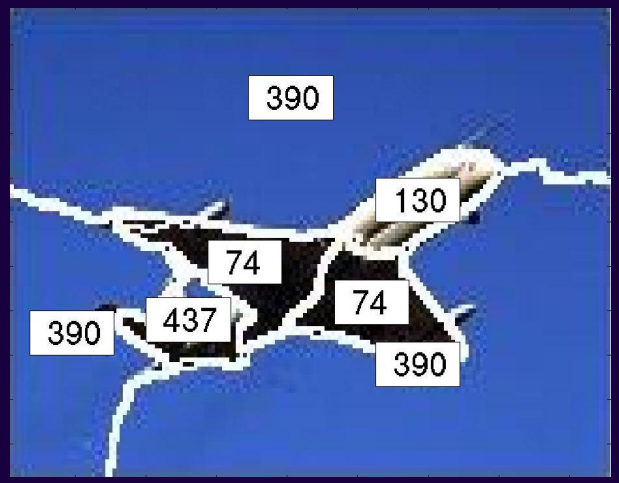
160 CD's
100 images in each

10 sets
each :
- randomly selected 80 CD's
- ~6000 training
- ~2000 test
- 150-200 word tokens
- 500 blob tokens

Segmentation
- about a month

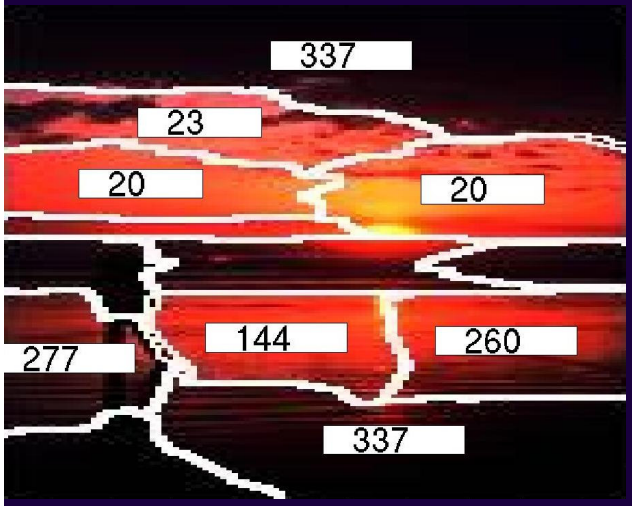


city mountain sky sun

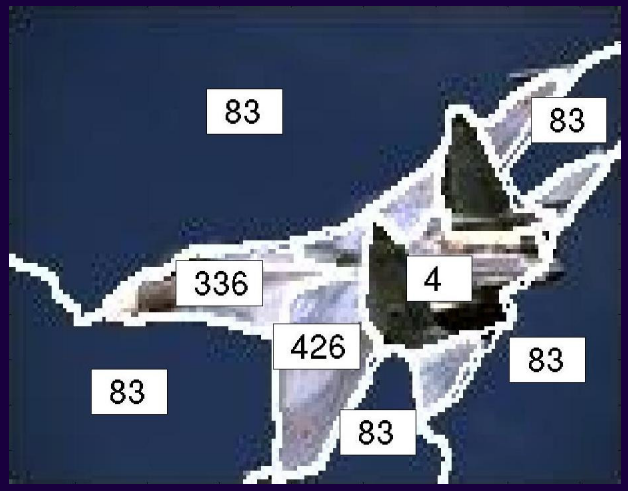


jet plane sky

cat forest grass tiger



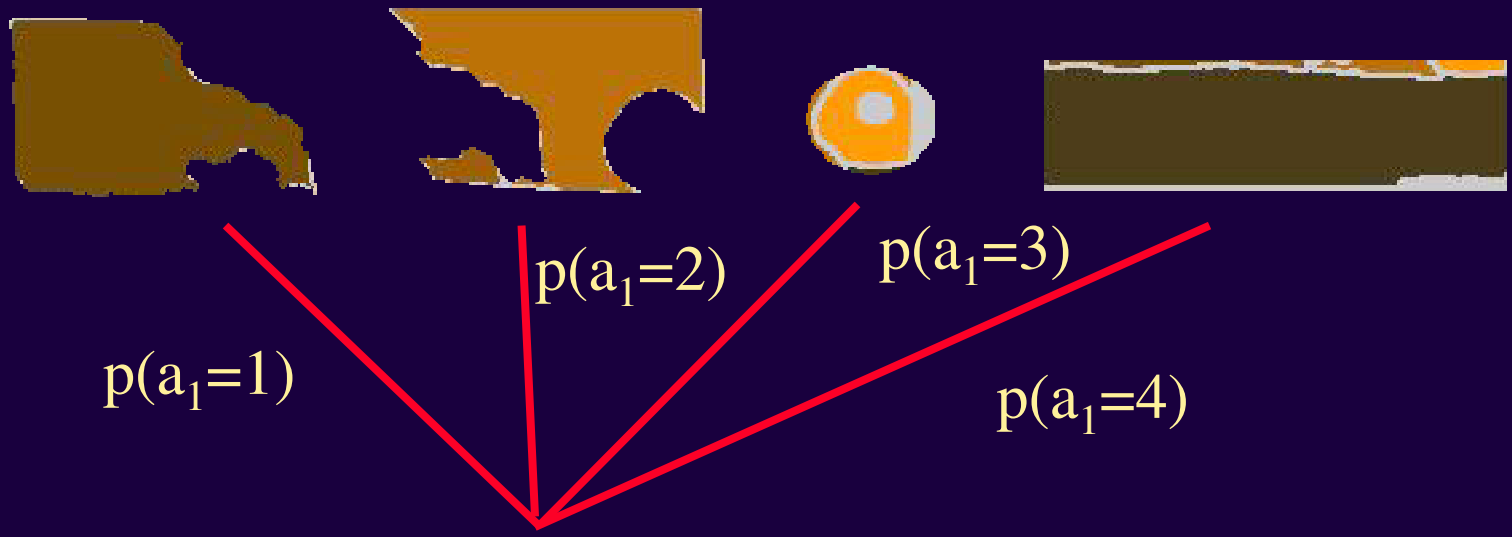
beach people sun water



jet plane sky

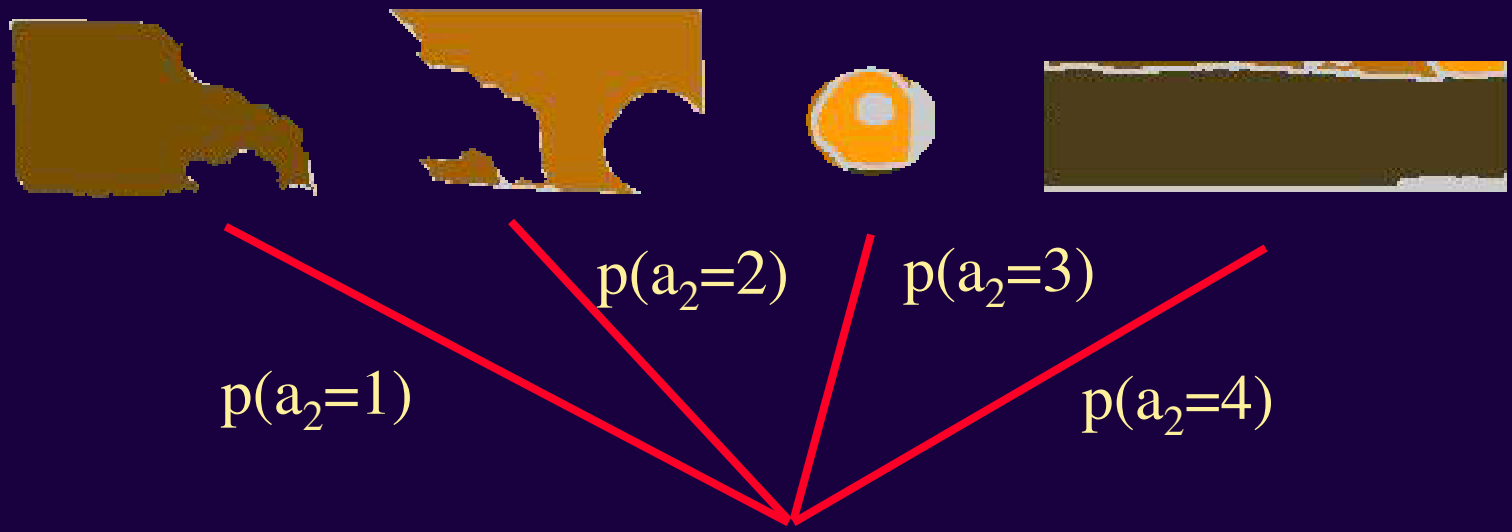
cat grass tiger water

Assignments



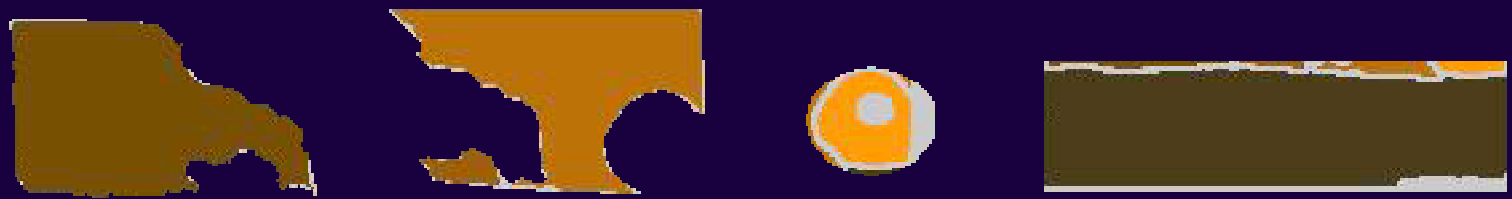
$$\sum_{i=1}^{B_n} p(a_1 = i) = 1$$

Assignments



$$\sum_{i=1}^{B_n} p(a_2 = i) = 1$$

Assignments



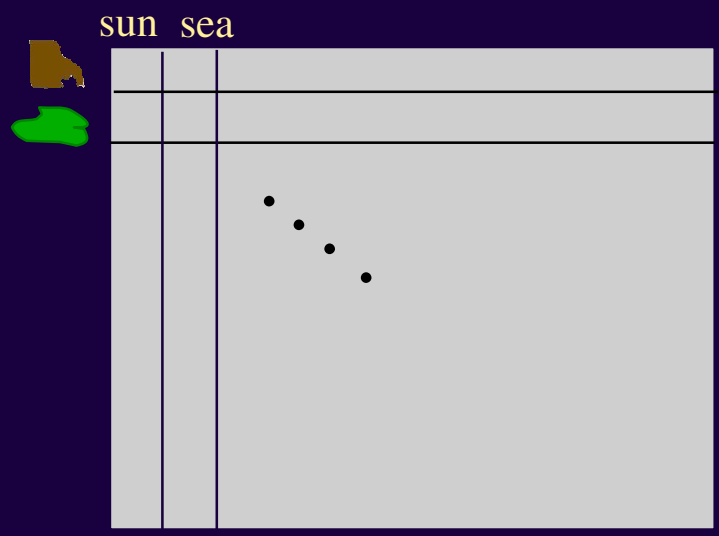
$p(a_3=1)$ $p(a_3=2)$ $p(a_3=3)$ $p(a_3=4)$

"sun sea sky"

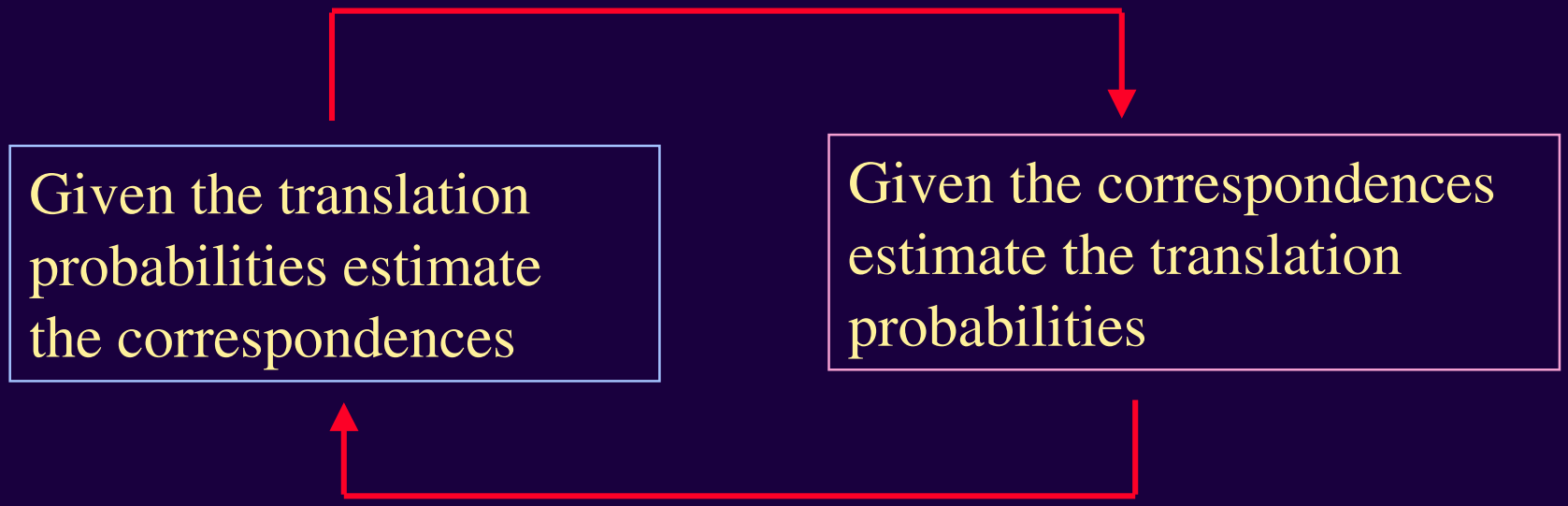
$$\sum_{i=1}^{B_n} p(a_3 = i) = 1$$

Initialization

Initialize translation table to blob-word cooccurences (emprical joint distribution of blobs and words)



Expectation Maximization

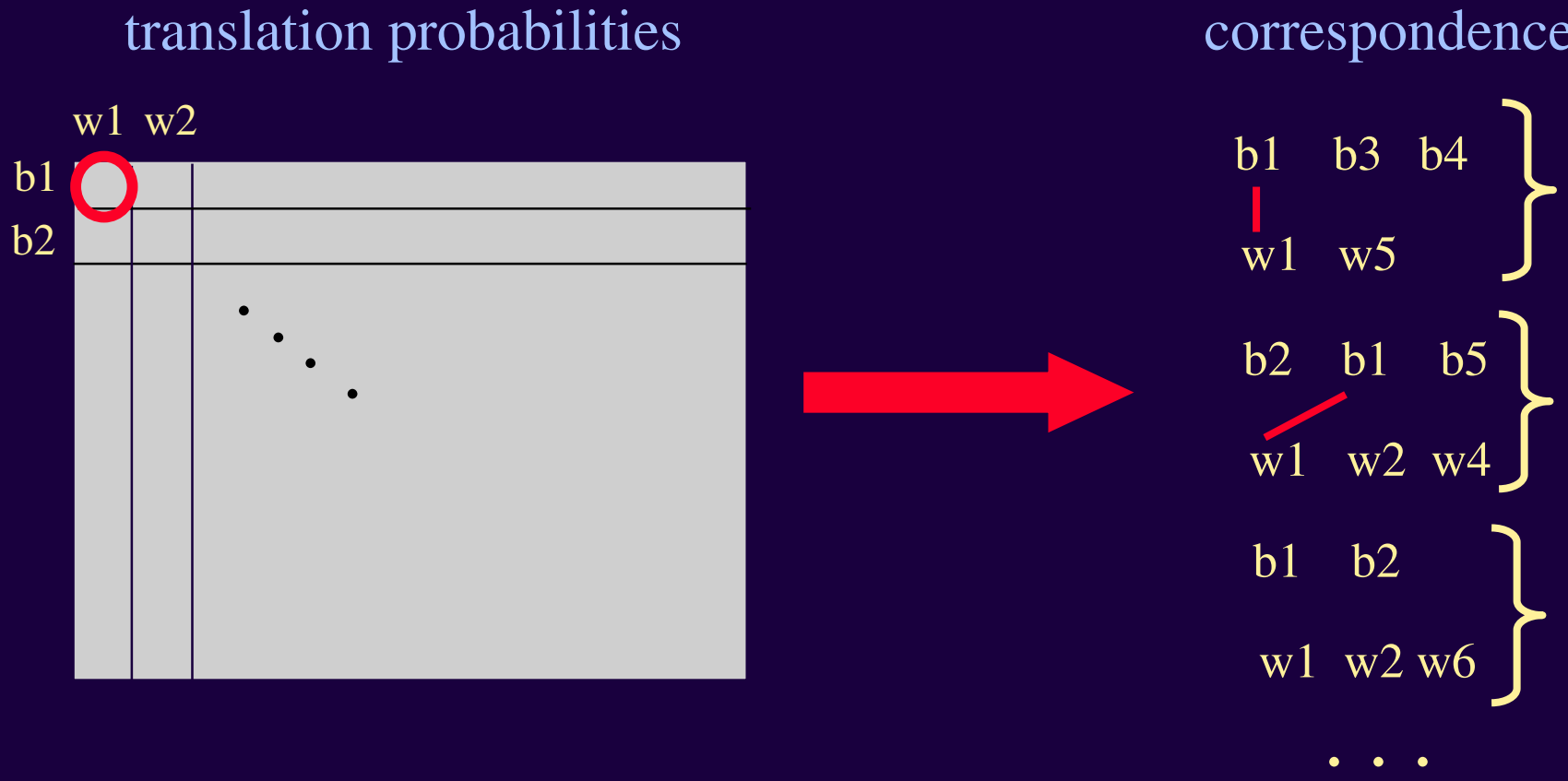


Dempster et al., 77

EM algorithm

E step : Predicting correspondences from translation probabilities

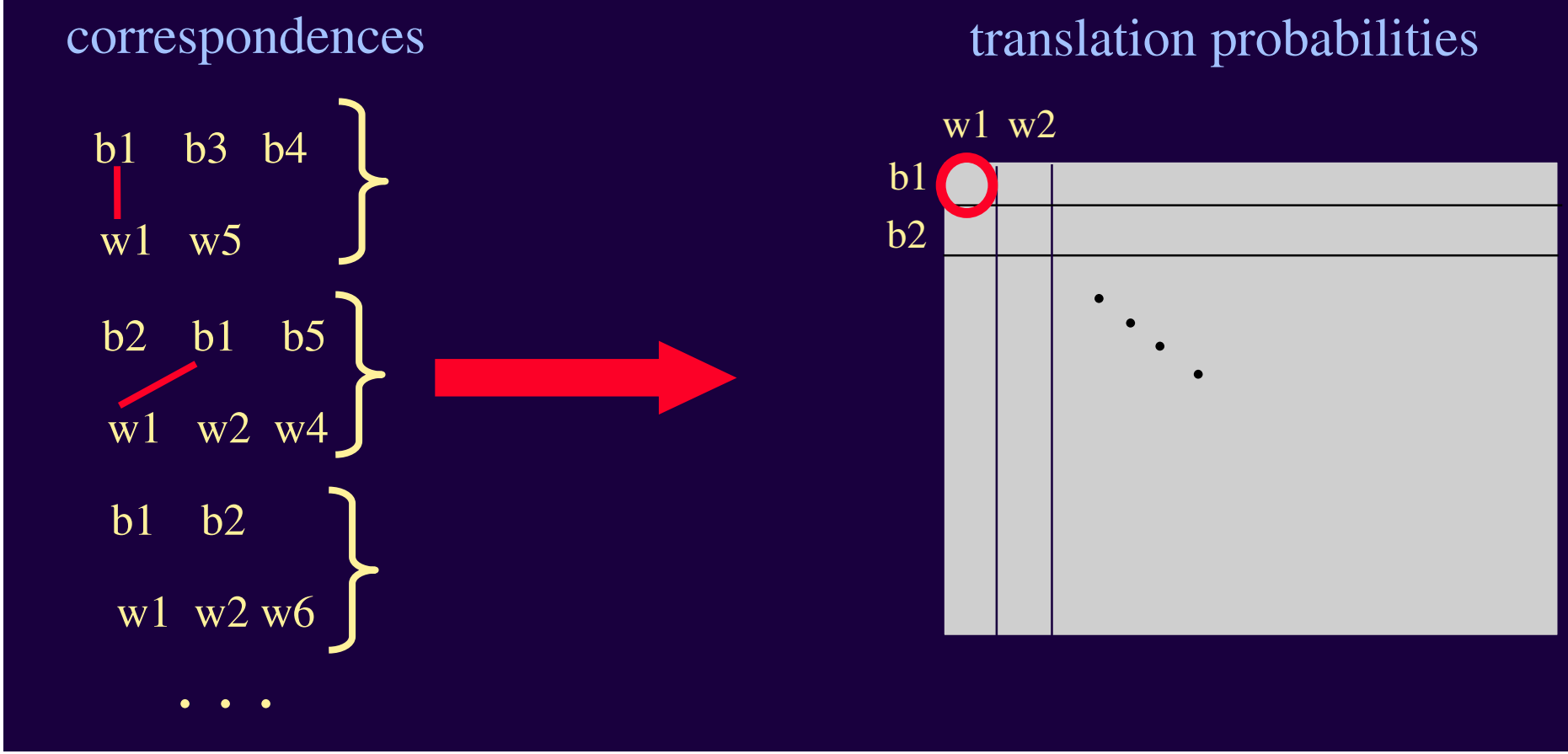
(for one pair)



EM algorithm

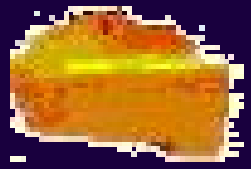
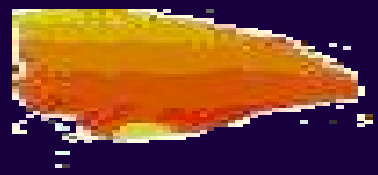
M step : Predicting translation probabilities from correspondences

(for one pair)



Dictionary

sun

sky



cat



horse

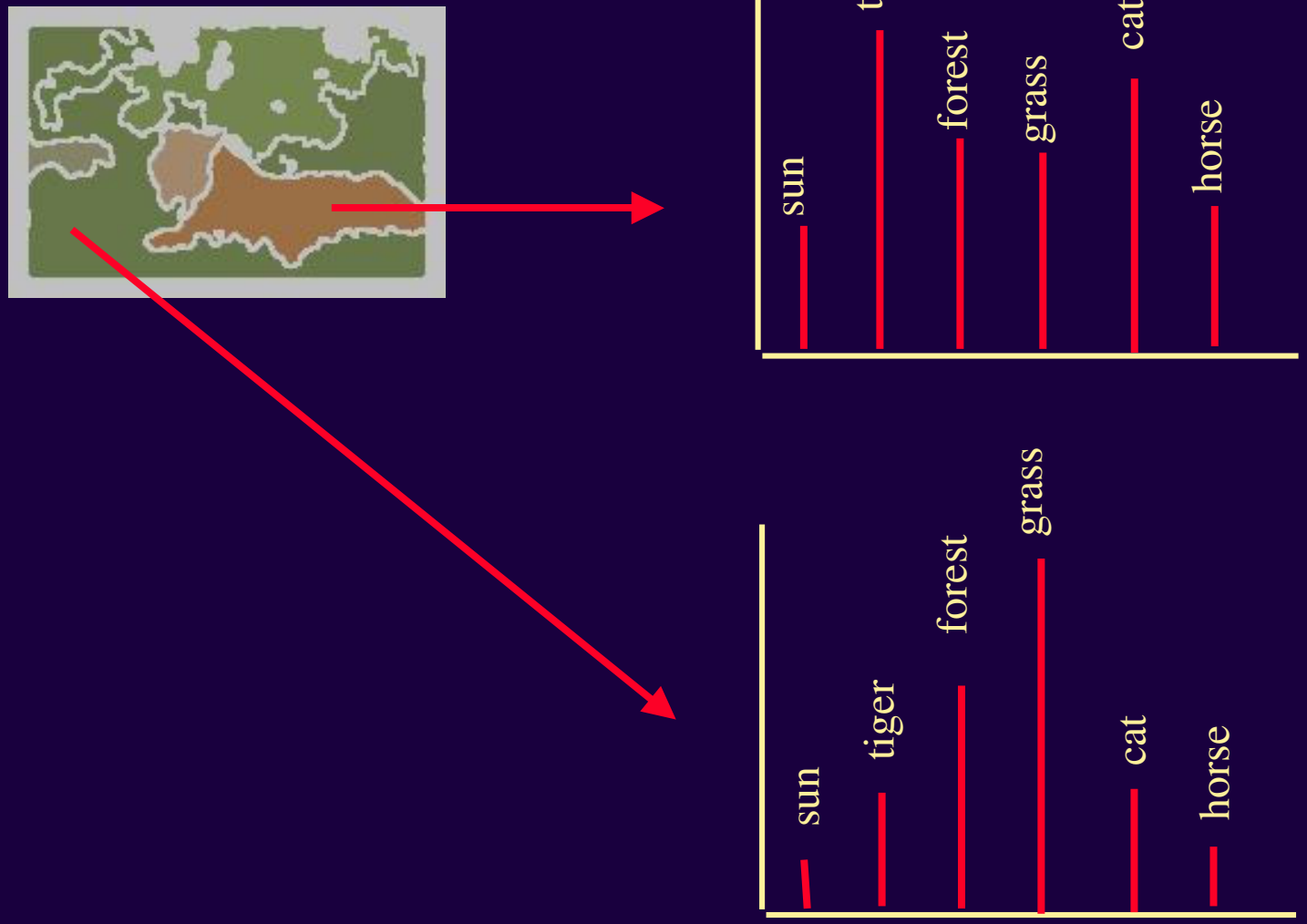


Labeling Regions

On a new image

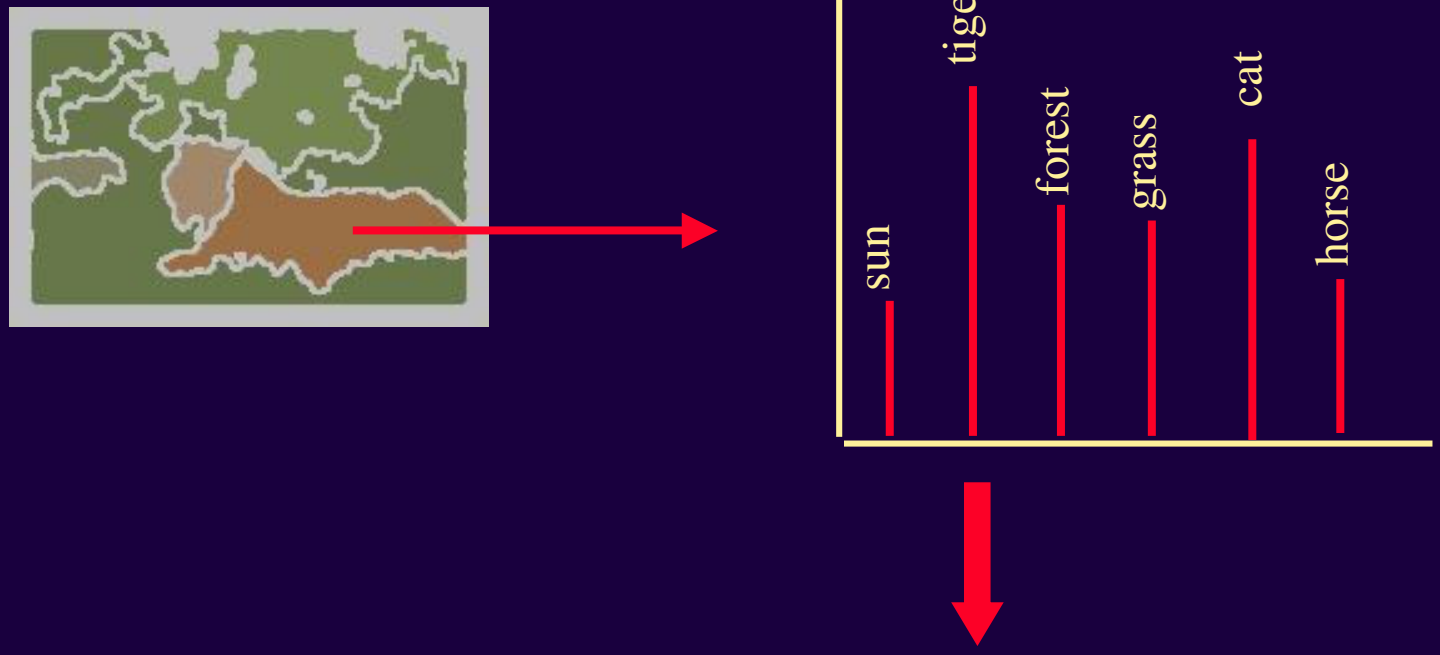
- Segment the image
- For each region
  - Find the blob token
  - Look at the word posterior given the blob

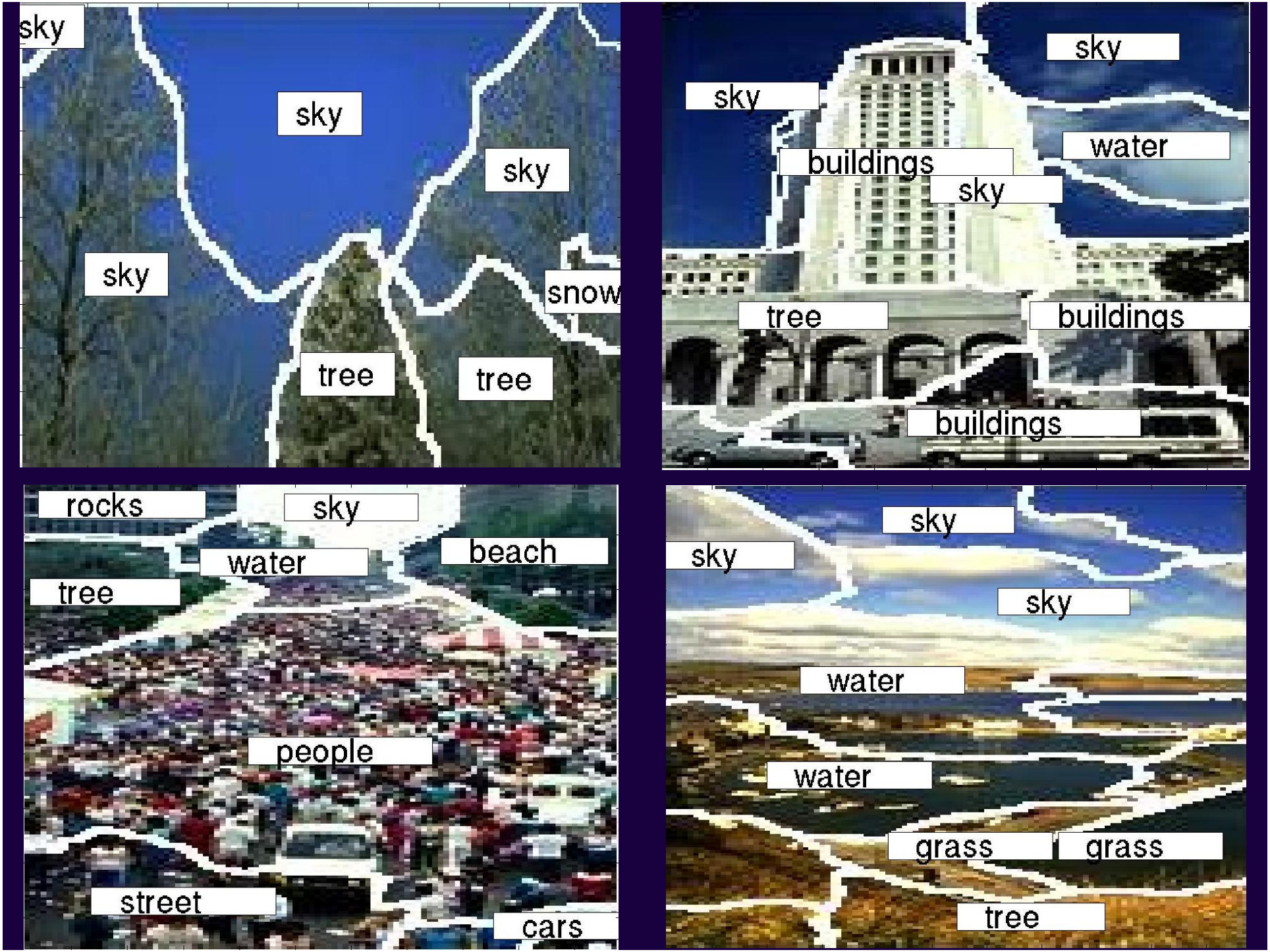
Labeling Regions

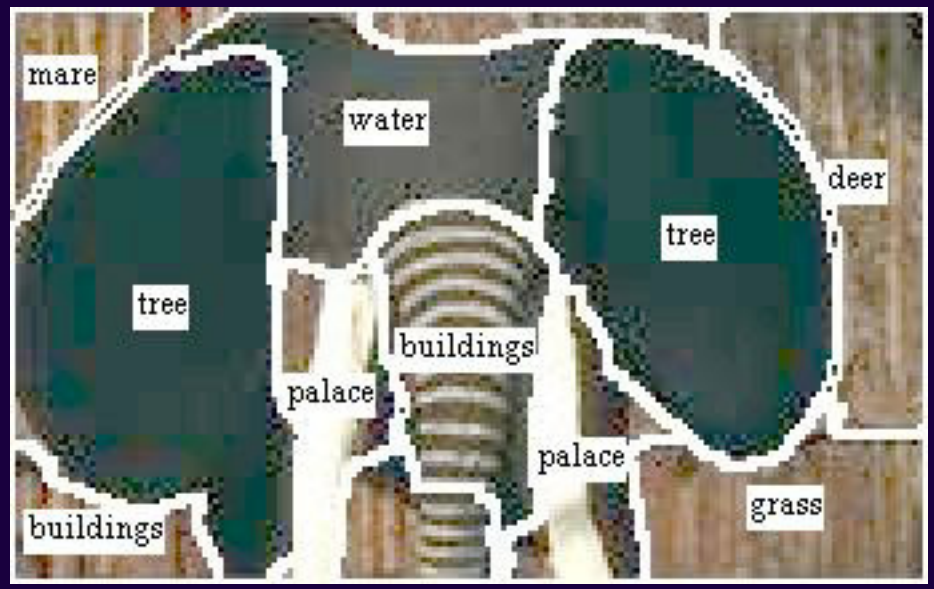


Labeling Regions

Display only maximal probable word




sky
sky
sky
sky
snow
tree
tree
sky
sky
sky
water
buildings
sky
tree
buildings
buildings
rocks
sky
water
beach
tree
people
street
cars
sky
sky
sky
water
water
grass
grass
tree


mare
water
deer
tree
tree
buildings
palace
palace
grass
buildings

Measuring Performance

First strategy--score by hand

Second strategy--use annotation performance as a proxy.

First Strategy
Score by hand

Average performance is four times better than guessing the most common word

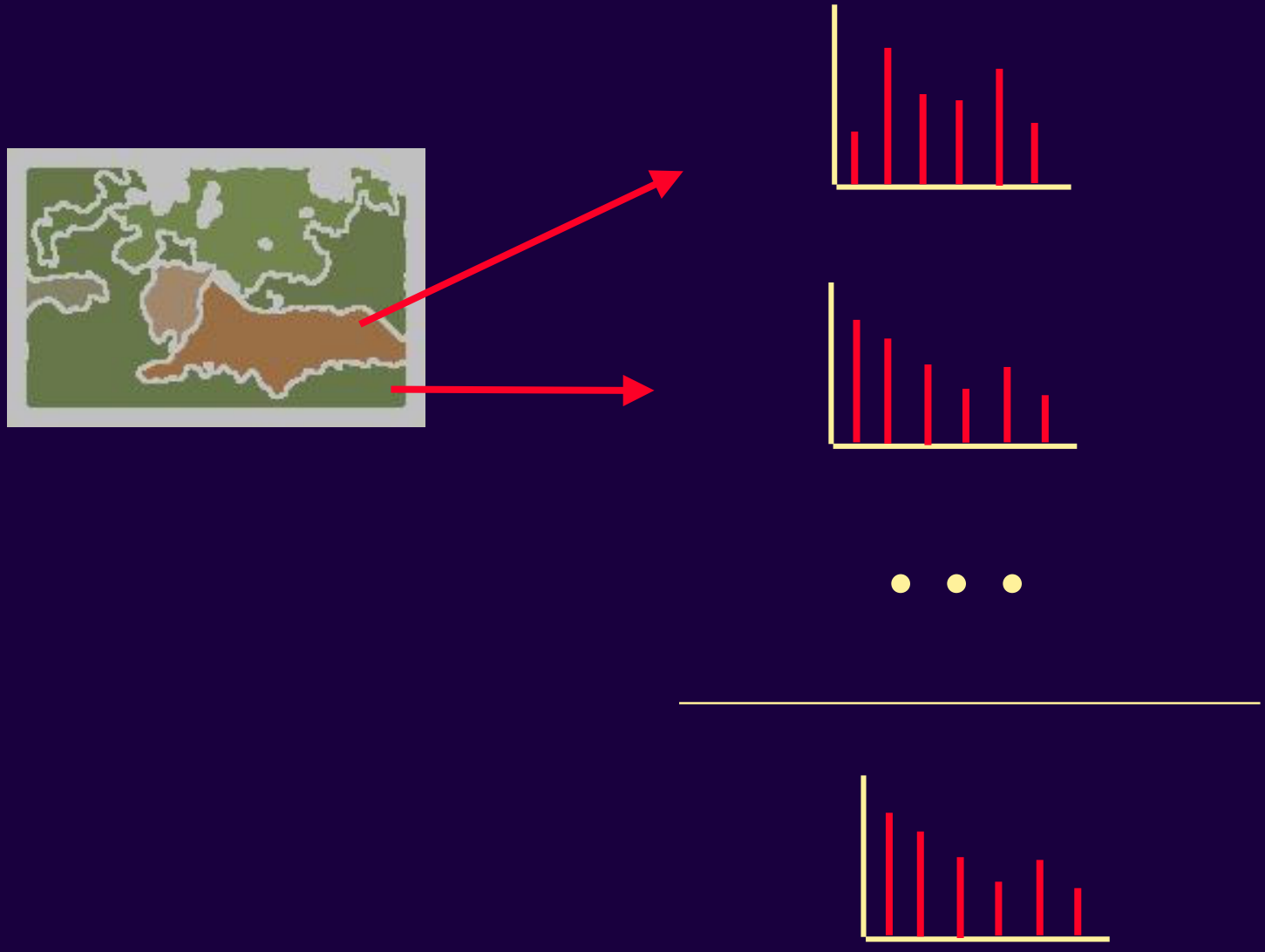
(“water”)

Second Strategy
Use Annotation

tiger cat grass water

Automatic : Don’t need to do by hand

Annotating Images



Measuring Annotation Performance

Actual Keywords → GRASS TIGER CAT FOREST

Predicted Words → CAT HORSE GRASS WATER

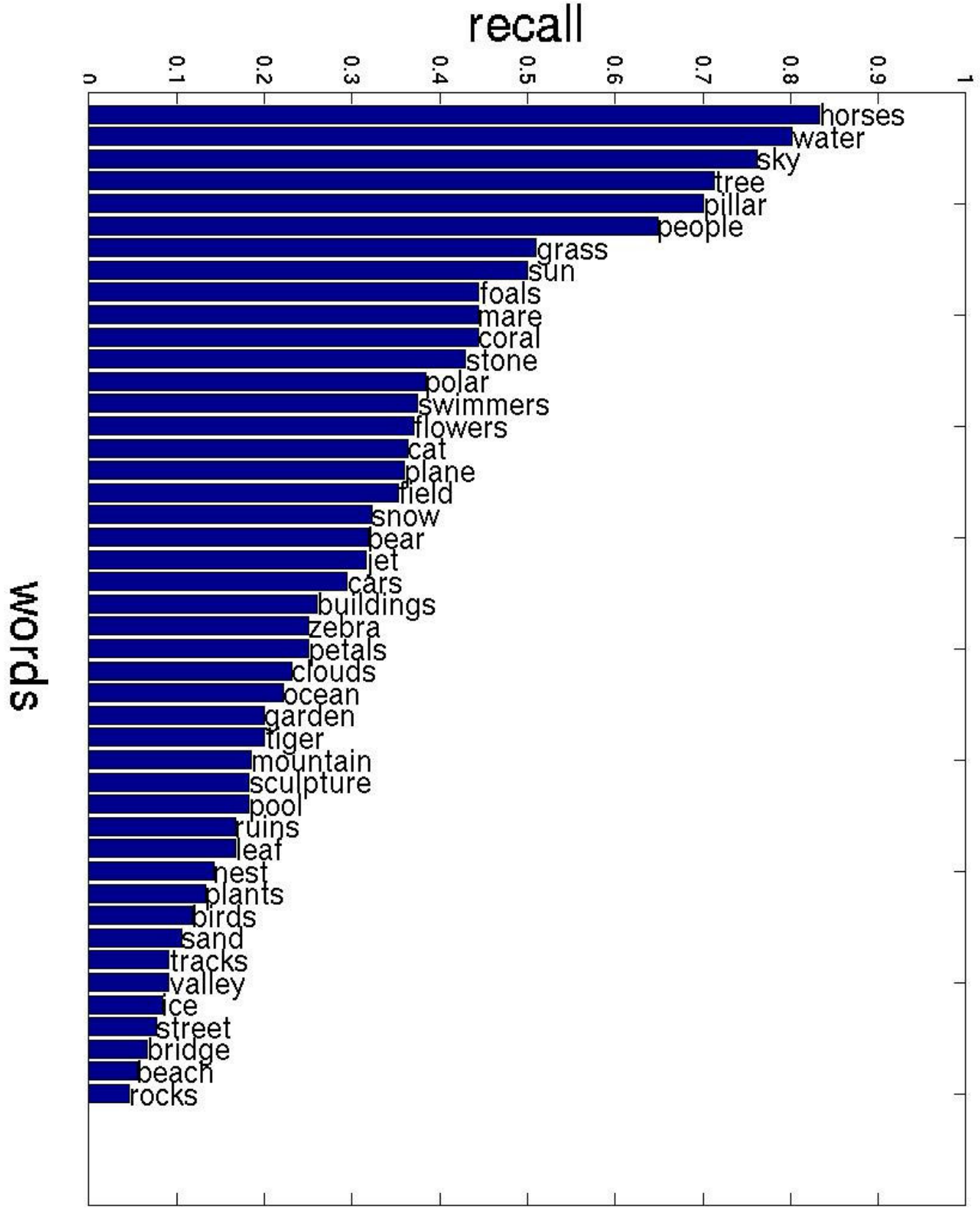
Measuring Annotation Performance

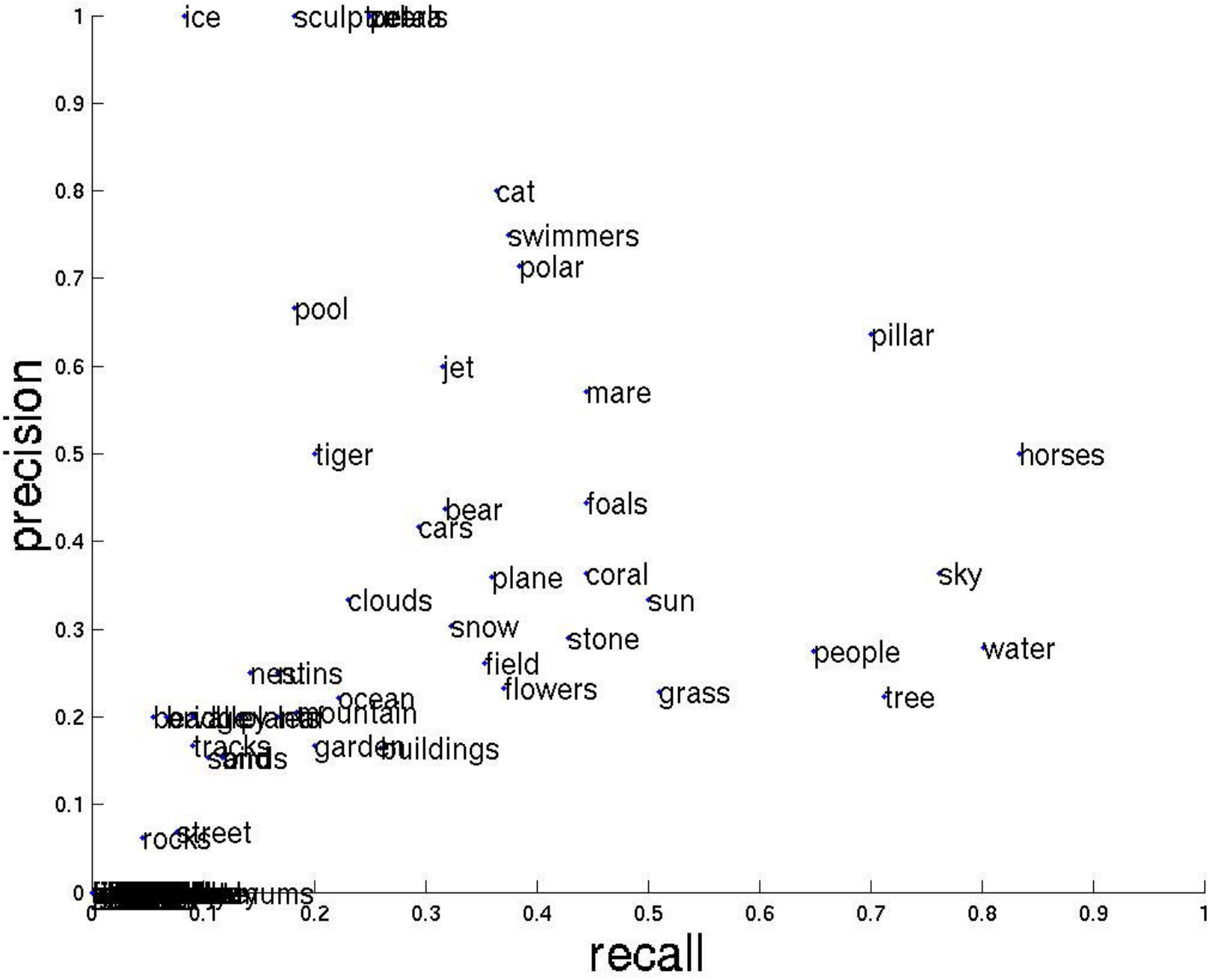
Actual Keywords

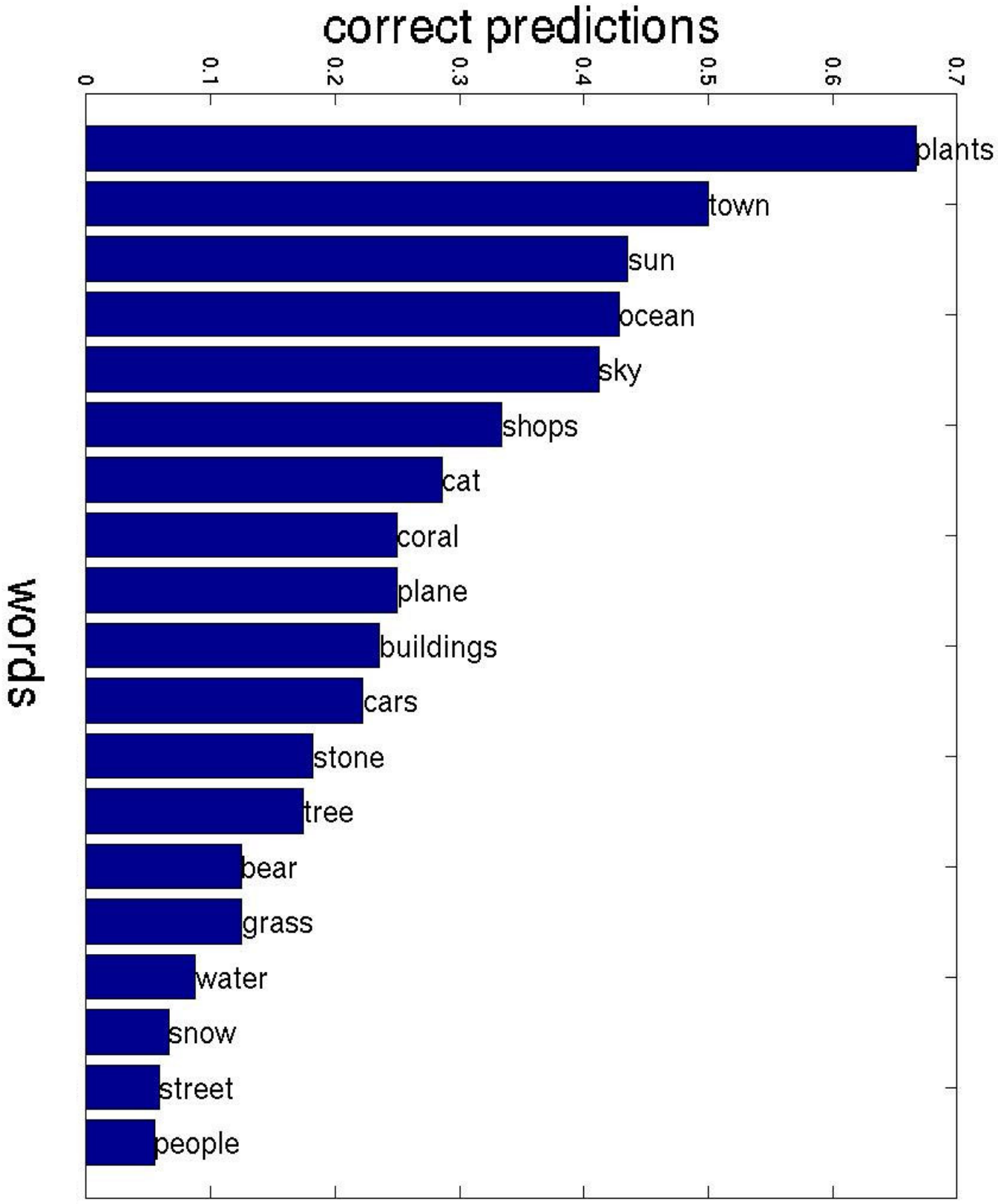
GRASS TIGER CAT FOREST

Predicted Words

CAT HORSE GRASS WATER


recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
words
horses
water
sky
tree
pillar
people
grass
sun
foals
mare
coral
stone
polar
swimmers
flowers
cat
plane
field
snow
bear
jet
cars
buildings
zebra
petals
clouds
ocean
garden
tiger
mountain
sculpture
pool
ruins
leaf
nest
plants
birds
sand
tracks
valley
ice
street
bridge
beach
rocks


precision
recall
ice
cat
swimmers
polar
pool
pillar
jet
mare
tiger
horses
bear
foals
cars
plane
coral
sky
clouds
sun
snow
stone
people
water
field
flowers
grass
tree
ocean
mountain
tracks
garden
buildings
sand
rocks
street
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


correct predictions
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
words
plants
town
sun
ocean
sky
shops
cat
coral
plane
buildings
cars
stone
tree
bear
grass
water
snow
street
people

Improving the System

- Refusing to predict
- Merging indistinguishable words

Refusing to predict

Null and fertility problems
simple solution to null - refusing to predict

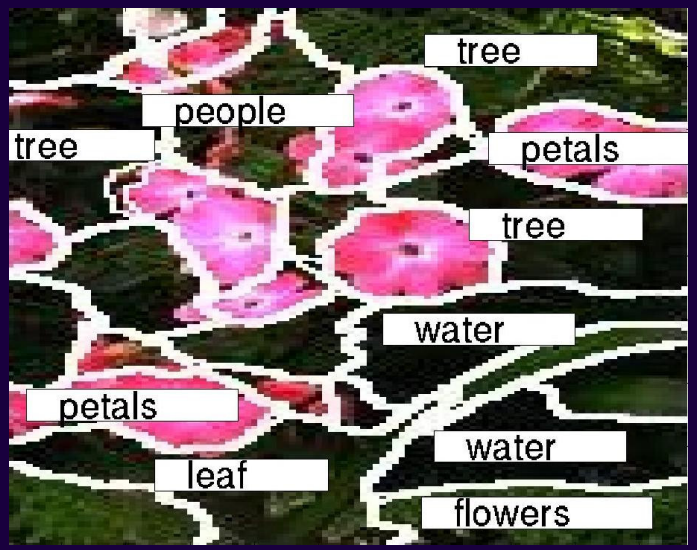
if p(word | blob) > threshold

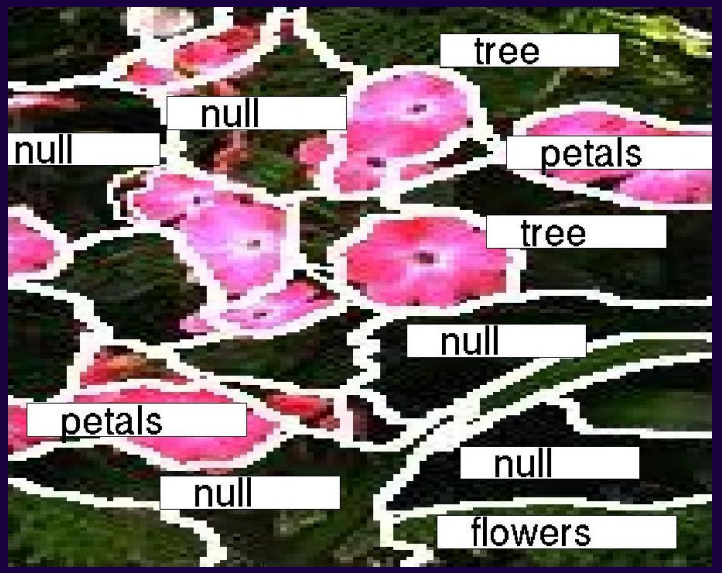
predict a word

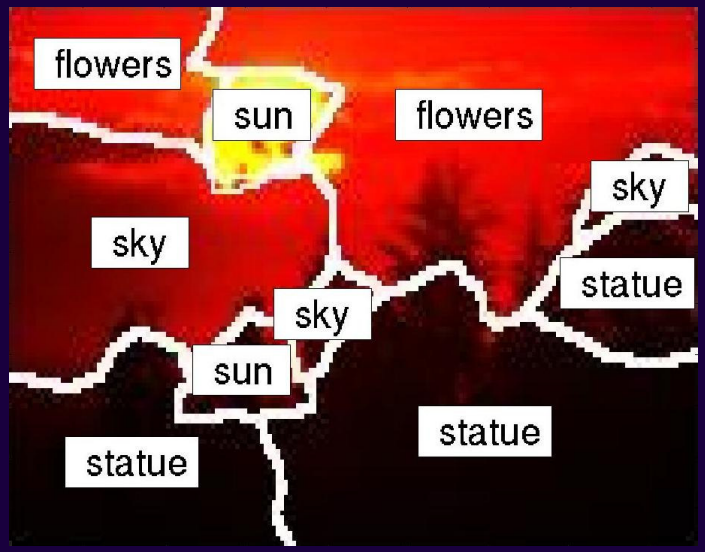
otherwise

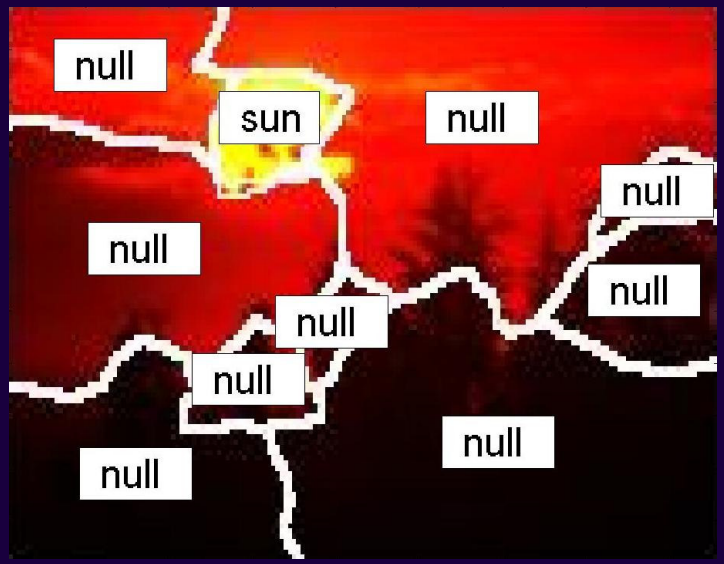
assign null

Examples
(null threshold = 0.2)

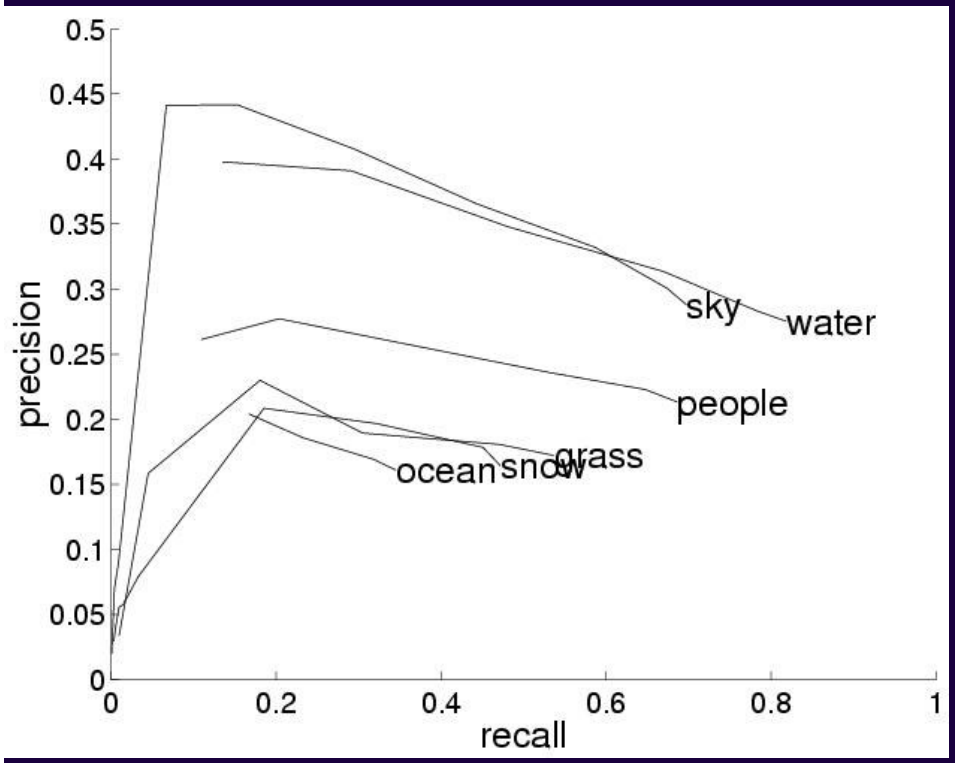




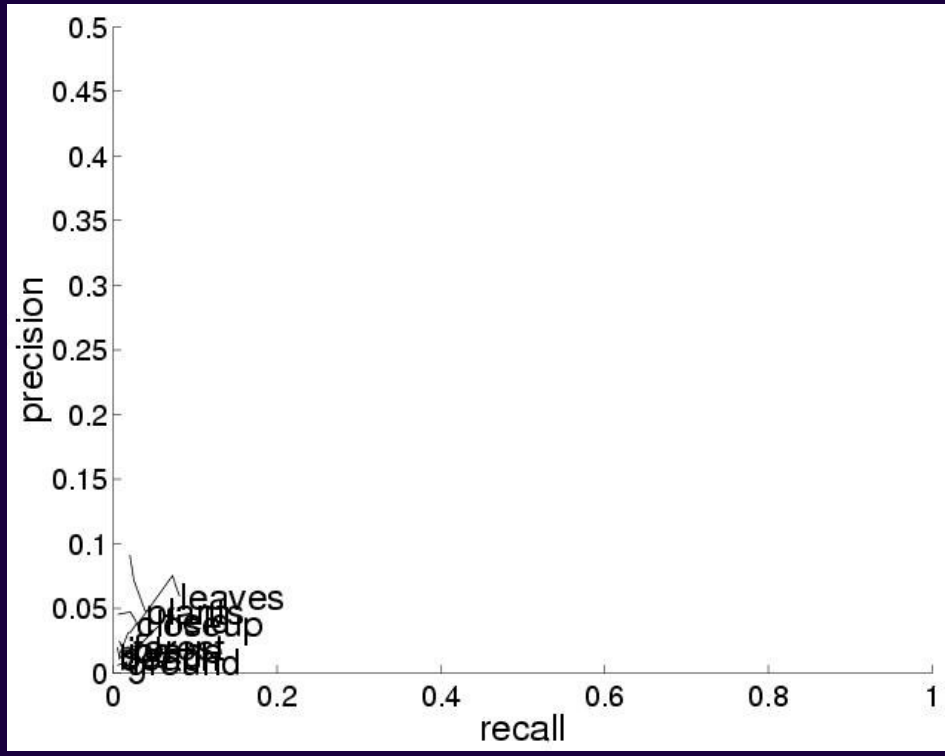




Recall and Precision
(for null threshold from 0 to 0.5)



selected good words



selected bad words

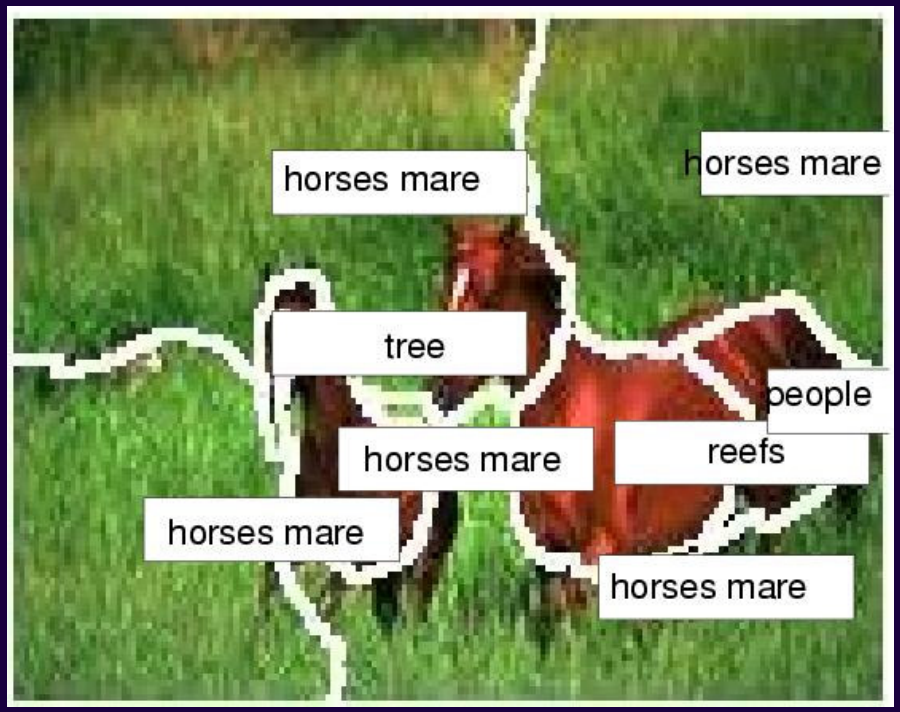
Clustering Indistinguishable Words

merge words which can't be told apart

e.g. locomotive vs. train

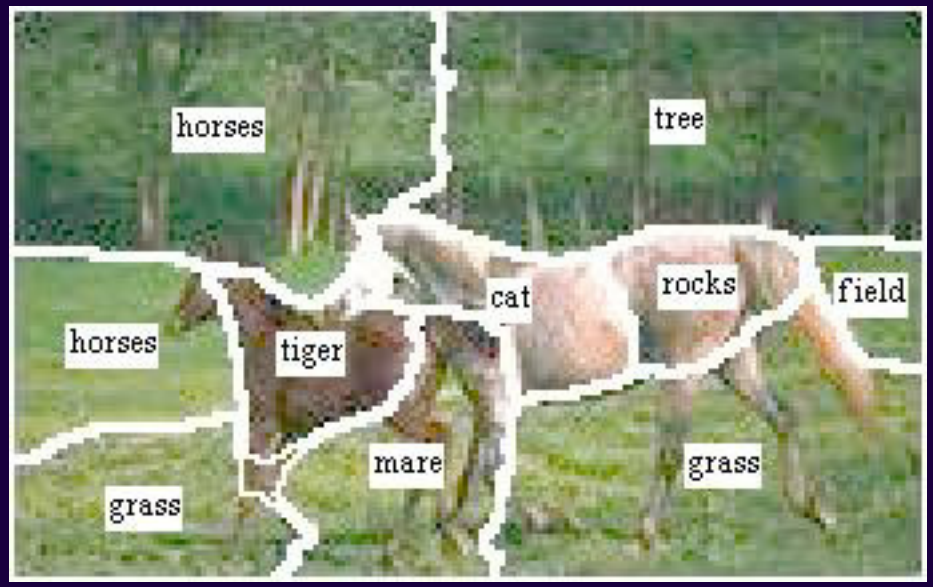
Examples


sky beach
train
train locomotive
train locomotive
tree
tree
reflection
tree
horses mare
stone


horses mare
horses mare
tree
people
horses mare
reefs
horses mare
horses mare

Future Directions
(machine learning)

Estimate where a minimal amount of supervision can be most helpful (and provide it)

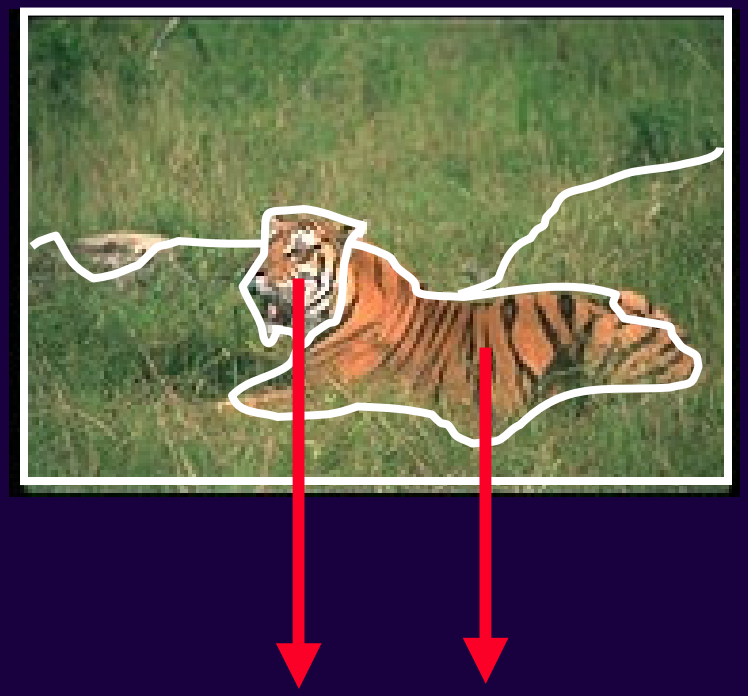


Future Directions
(computer vision)

Propose good features to differentiate words that are not distinguishable (e.g., eagle and jet)

Future Directions
(computer vision)

Propose region merging based on posterior word probabilities



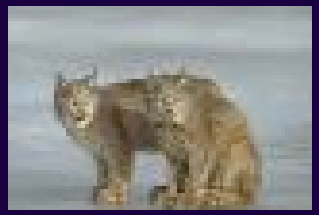
Propose merging

Conclusions

Recognition on the large scale

Unsupervised - using the available data efficiently

Learn what to recognize



The End